An information-theoretic derivation of min-cut based clustering

Anil Raj
Department of Applied Physics and Applied Mathematics
*Columbia University, New York**

Chris H. Wiggins
Department of Applied Physics and Applied Mathematics
Center for Computational Biology and Bioinformatics
Columbia University, New York†

(Dated: November 26, 2008)

Min-cut clustering, based on minimizing one of two heuristic cost-functions proposed by Shi and Malik, has spawned tremendous research, both analytic and algorithmic, in the graph partitioning and image segmentation communities over the last decade. It is however unclear if these heuristics can be derived from a more general principle facilitating generalization to new problem settings. Motivated by an existing graph partitioning framework, we derive relationships between optimizing relevance information, as defined in the Information Bottleneck method, and the regularized cut in a K-partitioned graph. For fast mixing graphs, we show that the cost functions introduced by Shi and Malik can be well approximated as the rate of loss of predictive information about the location of random walkers on the graph. For graphs generated from a stochastic algorithm designed to model community structure, the optimal information theoretic partition and the optimal min-cut partition are shown to be the same with high probability.

Keywords: graphs, clustering, information theory, min-cut, information bottleneck, graph diffusion

1. INTRODUCTION

Min-cut based graph partitioning has been used successfully to find clusters in networks, with applications in image segmentation as well as clustering biological and sociological networks. The central idea is to develop fast and efficient algorithms that optimally cut the edges between graph nodes, resulting in a separation of graph nodes into clusters. Particularly, since Shi and Malik successfully showed [1] that the *average* cut and the *normalized* cut (defined below) were useful heuristics to be optimized, there has been tremendous research in constructing the best normalized-cut-based cost function in the image segmentation community.

The Information Bottleneck (IB) method [2, 3] is a clustering technique, based on rate-distortion theory [4], that has been successfully applied in a wide variety of contexts including clustering word documents and gene-expression profiles [5]. The IB method is also capable of learning clusters in graphs and has been used successfully for synthetic and actual networks [6]. In the hard clustering case, given the diffusive probability distribution over a graph, IB optimally assigns probability distributions, associated with nodes, into distinct groups. These assignment rules define a separation of the graph nodes into clusters.

We here illustrate how minimizing the two cut-based heuristics introduced by Shi and Malik can be well-approximated by the rate of loss of *relevance information*, defined in the IB method applied to clustering graphs. To establish these relations, we must first define the graphs to be partitioned; we assume hard-clustering and the cluster cardinality to be K. We show, numerically, that maximizing mutual information and minimizing *regularized* cut amount to the same partition with high probability, for more modular 32-node graphs, where *modularity* is defined by the probability of inter-cluster edge connections in the Stochastic Block Model for graphs (See Numerical Experiments). We also show that the optimization goal of maximizing relevance information is equivalent to minimizing the regularized cut for 16-node graphs.[12]

2. THE MIN-CUT PROBLEM

Following [7], for an undirected, unweighted graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with n nodes and m edges, represented[13] by its adjacency matrix $\mathbf{A} := \{A_{xy} = 1 \iff x \sim y\}$, we define for two not necessarily disjoint sets of nodes $\mathbf{V}_+, \mathbf{V}_- \subseteq \mathbf{V}$, the association

$$W(\mathbf{V}_+, \mathbf{V}_-) = \sum_{x \in \mathbf{V}_+, y \in \mathbf{V}_-} A_{xy}. \tag{2.1}$$

We define a bisection of $\mathbf{V}$ into $\mathbf{V}_\pm$ if $\mathbf{V}_+ \cup \mathbf{V}_- = \mathbf{V}$ and $\mathbf{V}_+ \cap \mathbf{V}_- = \emptyset$. For a bisection of $\mathbf{V}$ into $\mathbf{V}_+$ and $\mathbf{V}_-$, the 'cut' is defined as $c(\mathbf{V}_+, \mathbf{V}_-) = W(\mathbf{V}_+, \mathbf{V}_-)$. We also quantify the size of a set $\mathbf{V}_+ \subseteq \mathbf{V}$ in terms of the number of nodes in the set $\mathbf{V}_+$ or the number of edges

*Electronic address: ar2384@columbia.edu
†Electronic address: chris.wiggins@columbia.edu

with at least one node in the set $\mathbf{V}_+$:

$$\begin{aligned}\omega(\mathbf{V}_+) &= \sum_{x\in\mathbf{V}_+} 1 \\ \Omega(\mathbf{V}_+) &= \sum_{x\in\mathbf{V}_+} d_x, \end{aligned} \tag{2.2}$$

where d_x is the degree of node x.

Shi and Malik [1] defined a pair of regularized cuts, for a bisection of $\mathbf{V}$ into $\mathbf{V}_+$ and $\mathbf{V}_-$; the *average cut* was defined as

$$\mathcal{A} = \frac{W(\mathbf{V}_+,\mathbf{V}_-)}{\omega(\mathbf{V}_+)} + \frac{W(\mathbf{V}_+,\mathbf{V}_-)}{\omega(\mathbf{V}_-)} \tag{2.3}$$

and the *normalized cut* was defined as

$$\mathcal{N} = \frac{W(\mathbf{V}_+,\mathbf{V}_-)}{\Omega(\mathbf{V}_+)} + \frac{W(\mathbf{V}_+,\mathbf{V}_-)}{\Omega(\mathbf{V}_-)}. \tag{2.4}$$

This definition can be generalized, for a K-partition of $\mathbf{V}$ into $\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_K$ [7], to

$$\mathcal{A} = \sum_j \frac{W(\mathbf{V}_j,\bar{\mathbf{V}}_j)}{\omega(\mathbf{V}_j)} \tag{2.5}$$

$$\mathcal{N} = \sum_j \frac{W(\mathbf{V}_j,\bar{\mathbf{V}}_j)}{\Omega(\mathbf{V}_j)} \tag{2.6}$$

where $\bar{\mathbf{V}}_j = \mathbf{V} \setminus \mathbf{V}_j$.

For the graph $\mathcal{G}$, we can define the graph Laplacian $\mathbf{\Delta} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is a diagonal matrix of vertex degrees. For a bisection of $\mathbf{V}$, we also define the partition indicator vector $\mathbf{h}$ [8]

$$h_x = \begin{cases} +1 & \forall x \in \mathbf{V}_+ \\ -1 & \forall x \in \mathbf{V}_-. \end{cases} \tag{2.7}$$

Specifying two 'prior' probability distributions over the set of nodes $\mathbf{V}$: (i) $p(x) \propto 1$ and (ii) $p(x) \propto d_x$, we then define the *average* of $\mathbf{h}$ to be

$$\begin{aligned}\bar{\mathbf{h}} &= \frac{\sum_{x\in\mathbf{V}} h_x}{n} \\ \langle\mathbf{h}\rangle &= \frac{\sum_{x\in\mathbf{V}} d_x h_x}{2m}. \end{aligned} \tag{2.8}$$

The cut, as defined by Fiedler [8], and the regularized cuts, as defined by Shi and Malik [1], can then by written in terms of $\mathbf{h}$ as (See Appendix)

$$\begin{aligned} c &= \frac{1}{4}\mathbf{h^T\Delta h} \\ \mathcal{A} &= \frac{1}{n}\frac{\mathbf{h^T\Delta h}}{1-\bar{\mathbf{h}}^2} \\ \mathcal{N} &= \frac{1}{2m}\frac{\mathbf{h^T\Delta h}}{1-\langle\mathbf{h}\rangle^2}. \end{aligned} \tag{2.9}$$

More generally, for a K-partition, we define the partition indicator matrix $\mathbf{Q}$ as

$$Q_{zx} \equiv p(z|x) = 1 \;\; \forall x \in z \tag{2.10}$$

where $z \in \{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_K\}$ and define $\mathbf{P}$ as a diagonal matrix of the 'prior' probability distribution over the nodes. The regularized cut can then be generalized as

$$\mathcal{C} = \sum_j \frac{[\mathbf{Q^T\Delta Q}]_{jj}}{[\mathbf{Q^TPQ}]_{jj}} \tag{2.11}$$

where for $p(x) \propto 1$, $\mathcal{C} = \mathcal{A}$; and for $p(x) \propto d_x$, $\mathcal{C} = \mathcal{N}$.

Inferring the optimal $\mathbf{h}$ (or $\mathbf{Q}$), however, has been shown to be an NP-hard combinatorial optimization problem [9].

3. INFORMATION BOTTLENECK

Rate-distortion theory, which provides the foundations for lossy data compression, formulates clustering in terms of a compression problem; it determines the code with minimum average length such that information can be transmitted without exceeding some specified distortion. Here, the model-complexity, or *rate*, is measured by the mutual information between the data and their representative codewords (average number of bits used to store a data point). Simpler models correspond to smaller rates but they typically suffer from relatively high *distortion*. The distortion measure, which can be identified with loss functions, usually depends on the problem; in the simplest of cases, it is the variance of the difference between data and their representatives.

The Information Bottleneck (IB) method [3] proposes the use of mutual information as a natural distortion measure. In this method, the data are compressed into clusters while maximizing the amount of information that the 'cluster representation' preserves about some specified *relevance* variable. For example, in clustering word documents, one could use the 'topic' of a document as the relevance variable.

For a graph $\mathcal{G}$, let X be a random variable over graph nodes, Y be the relevance variable and Z be the random variable over clusters. Graph partitioning using the IB method [6] learns a probabilistic cluster assignment function $p(z|x)$ which gives the probability that a given node x belongs to cluster z. The optimal $p(z|x)$ minimizes the mutual information between X and Z, while minimizing the loss of predictive information between Z and Y. This complexity–fidelity trade-off can be expressed in terms of a functional to be minimized

$$\mathcal{F}[p(z|x)] = -I[Y;Z] + TI[X;Z] \tag{3.1}$$

where the temperature T parameterizes the relative importance of precision over complexity. As $T \to 0$, we reach the 'hard clustering' limit where each node is assigned with unit probability to one cluster (i.e $p(z|x) \in \{0,1\}$).

Graph clustering, as formulated in terms of the IB method, requires a joint distribution $p(y, x)$ to be defined on the graph; we use the distribution given by continuous graph diffusion as it naturally captures topological information about the network [6]. The relevance variable Y then ranges over the nodes of the graph and is defined as the node at which a random walker ends at time t if the random walker starts at node x at time 0. For continuous time diffusion, the conditional distribution $p^t(y|x)$ is given as

$$\mathbf{G}^t = p^t(y|x) = e^{-t\mathbf{\Delta}\mathbf{P}^{-1}} \tag{3.2}$$

where $\mathbf{\Delta}$ is the graph Laplacian and $\mathbf{P}$ a diagonal matrix of the prior distribution over the graph nodes, as described earlier. The characteristic diffusion time scale τ of the system is given by the inverse of the smallest non-zero eigenvalue of the diffusion operator exponent $\mathbf{\Delta}\mathbf{P}^{-1}$ and characterizes the slowest decaying mode in the system. To calculate the joint distribution $p(y, x)$ from the conditional $\mathbf{G}^t$, we must specify an initial or prior distribution[14]; we use the two different priors $p(x)$, used earlier to calculate the expected value of $\mathbf{h}$: (i) $p(x) \propto 1$ and (ii) $p(x) \propto d_x$.

4. RATE OF INFORMATION LOSS IN GRAPH DIFFUSION

We analyze here the rate of loss of predictive information between the relevance variable Y and the cluster variable Z, during diffusion on a graph $\mathcal{G}$, after the graph nodes have been hard-partitioned into K clusters.

A. Well-mixed limit of graph diffusion

For a given partition $\mathbf{Q}$ of the graph, defined in Eqn. (2.10), we approximate the mutual information $I[Y;Z]$ when diffusion on the graph reaches its well-mixed limit. We introduce the *dependence* $\eta(y, z)$ such that

$$p(y, z) = p(y)p(z)(1 + \eta). \tag{4.1}$$

This implies $\langle\eta\rangle_y = \langle\eta\rangle_z = 0$ and $\left\langle\left\langle\eta^2\right\rangle_z\right\rangle_y = \langle\eta\rangle$ where $\langle\rangle$ denotes expectation over the joint distribution and $\langle\rangle_y$ and $\langle\rangle_z$ denote expectation over the corresponding marginals.

In the well-mixed limit, we have $\eta \ll 1$. The predictive information (expressed in nats) can then be approximated as:

$$\begin{aligned}
I[Y;Z] &= \left\langle \ln \frac{p(z,y)}{p(z)p(y)} \right\rangle \\
&= \left\langle \langle (1+\eta)\ln(1+\eta)\rangle_y \right\rangle_z \\
&\approx \left\langle \left\langle (1+\eta)(\eta - \frac{1}{2}\eta^2)\right\rangle_y \right\rangle_z \\
&\approx \left\langle \left\langle \eta + \frac{1}{2}\eta^2 \right\rangle_y \right\rangle_z \\
&= \frac{1}{2}\left\langle \langle \eta^2 \rangle_y \right\rangle_z \qquad (4.2) \\
&= \frac{1}{2}\sum_{y,z} p(y)p(z)\left(\frac{p(z,y)}{p(z)p(y)} - 1\right)^2 \\
&= \frac{1}{2}\left(\sum_{y,z}\frac{p(y,z)^2}{p(y)p(z)} - 1\right) \equiv \iota. \qquad (4.3)
\end{aligned}$$

Here, we define ι as a first-order approximation to $I[Y;Z]$ in the well-mixed limit of graph diffusion.

1. *Well-mixed K-partitioned graph*

As in the IB method, the Markov condition $Z - X - Y$ allows us to make several simplifications for the conditional distributions and associated information theoretic measures. For a K-partition $\mathbf{Q}$ of the graph, we have

$$\begin{aligned}
p(y,z) &= \sum_x p(x,y,z) \\
&= \sum_x p(z|y,x)p(y|x)p(x) \\
&= \sum_x p(z|x)p(y|x)p(x) \equiv \mathbf{Q}\mathbf{P}\mathbf{G}^{t\,\mathbf{T}}. \qquad (4.4) \\
p(y,z)^2 &= \left(\sum_x p(z|x)p(y|x)p(x)\right)^2 \\
&= \sum_{x,x'=1}^{n} p(z|x)p(y|x)p(x)p(z|x')p(y|x')p(x') \\
&= \sum_{x,x'=1}^{n} Q_{zx}G^t_{yx}P_x Q_{zx'}G^t_{yx'}P_{x'}. \qquad (4.5) \\
p(z) &= \sum_x p(z|x)p(x) \\
&= \sum_x Q_{zx}P_x. \qquad (4.6)
\end{aligned}$$

Graph diffusion being a Markov process, we have $\sum_{y=1}^{n} G^t_{x'y}G^t_{yx} = G^{2t}_{x'x}$. Using this and Bayes rule

$G^t_{yx}P_x = G^t_{xy}P_y$, we have

$$\begin{aligned}
\iota &= \frac{1}{2}\left(\sum_{y,z}\frac{\sum_{x,x'=1}^{n} Q_{zx}G^t_{yx}P_xQ_{zx'}G^t_{yx'}P_{x'}}{(\sum_{x''}Q_{zx''}P_{x''})P_y} - 1\right)\\
&= \frac{1}{2}\left(\sum_{y,z}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}P_yG^t_{x'y}G^t_{yx}P_x}{(\sum_{x''}Q_{zx''}P_{x''})P_y} - 1\right)\\
&= \frac{1}{2}\left(\sum_{z=1}^{K}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}(\sum_{y=1}^{n}G^t_{x'y}G^t_{yx})P_x}{(\sum_{x''}Q_{zx''}P_{x''})} - 1\right)\\
&= \frac{1}{2}\left(\sum_{z=1}^{K}\frac{\sum_{x,x'=1}^{n} Q_{zx}Q_{zx'}G^{2t}_{x'x}P_x}{(\sum_{x''}Q_{zx''}P_{x''})} - 1\right). \qquad (4.7)
\end{aligned}$$

In the hard clustering case, $\sum_x Q_{zx}P_x = p(z) = [\mathbf{QPQ^T}]_{zz}$ and we have

$$\iota = \frac{1}{2}\left(\sum_{z=1}^{K}\frac{[\mathbf{Q}(\mathbf{G}^{2t}\mathbf{P})\mathbf{Q^T}]_{zz}}{[\mathbf{QPQ^T}]_{zz}} - 1\right). \qquad (4.8)$$

2. Well-mixed 2-partitioned graph

We can re-write ι as

$$\begin{aligned}
\iota &= \frac{1}{2}\left\langle\left\langle\eta^2\right\rangle_y\right\rangle_z\\
&= \frac{1}{2}\left\langle\left\langle\frac{(p(z|y)-p(z))^2}{p(z)^2}\right\rangle_z\right\rangle_y. \qquad (4.9)
\end{aligned}$$

For a bisection $\mathbf{h}$ of the graph, $z \in \{+1,-1\}$ and we have

$$\begin{aligned}
p(z|x) &= \frac{1}{2}(1 \pm h_x) \equiv \frac{1}{2}(1+zh_x). \qquad (4.10)\\
p(z|y) &= \frac{1}{p(y)}\sum_x p(z,y,x)\\
&= \frac{1}{p(y)}\sum_x p(z|x)p(y|x)p(x)\\
&= \frac{1}{2}\sum_x (1+zh_x)p(x|y)\\
&= \frac{1}{2}(1+z\langle\mathbf{h}|y\rangle). \qquad (4.11)\\
p(z) &= \sum_x p(z,x) = \sum_x p(z|x)p(x)\\
&= \frac{1}{2}\sum_x (1+zh_x)p(x)\\
&= \frac{1}{2}(1+z\langle\mathbf{h}\rangle). \qquad (4.12)\\
p(z|y)-p(z) &= \frac{1}{2}(1+z\langle\mathbf{h}|y\rangle) - \frac{1}{2}(1+z\langle\mathbf{h}\rangle)\\
&= \frac{1}{2}z(\langle\mathbf{h}|y\rangle - \langle\mathbf{h}\rangle). \qquad (4.13)
\end{aligned}$$

We then have

$$\begin{aligned}
\left\langle\frac{(p(z|y)-p(z))^2}{p(z)^2}\right\rangle_z &= \sum_{z=1}^{K}\frac{\frac{1}{4}(\langle\mathbf{h}|y\rangle - \langle\mathbf{h}\rangle)^2}{\frac{1}{2}(1+z\langle\mathbf{h}\rangle)}\\
&= \frac{(\langle\mathbf{h}|y\rangle - \langle\mathbf{h}\rangle)^2}{2}\sum_{z=1}^{K}\frac{1}{1+z\langle\mathbf{h}\rangle}\\
&= \frac{(\langle\mathbf{h}|y\rangle - \langle\mathbf{h}\rangle)^2}{1-\langle\mathbf{h}\rangle^2}. \qquad (4.14)
\end{aligned}$$

The mutual information $I[Y;Z]$ can then be approximated as

$$\begin{aligned}
\iota &= \frac{1}{2}\frac{\left\langle(\langle\mathbf{h}|y\rangle - \langle\mathbf{h}\rangle)^2\right\rangle_y}{1-\langle\mathbf{h}\rangle^2}\\
&= \frac{1}{2}\frac{\sigma^2_y(\langle\mathbf{h}|y\rangle)}{1-\langle\mathbf{h}\rangle^2}. \qquad (4.15)
\end{aligned}$$

Using Bayes rule $p^t(x|y)p(y) = p^t(y|x)p(x)$, we have

$$\begin{aligned}
\langle\mathbf{h}|y\rangle &= \sum_x h_x p^t(x|y) = \sum_x \frac{h_x p^t(y|x)p(x)}{p(y)}. \qquad (4.16)\\
\left\langle\langle\mathbf{h}|y\rangle^2\right\rangle_y &= \sum_{y=1}^{n} p(y)\sum_{x,x'=1}^{n} h_xh_{x'}\frac{p^t(y|x)p(x)p^t(x'|y)}{p(y)}\\
&= \sum_{y=1}^{n}\sum_{x,x'=1}^{n} h_xh_{x'}p^t(y|x)p^t(x'|y)p(x). \qquad (4.17)
\end{aligned}$$

Again, graph diffusion being a Markov process,

$$\begin{aligned}
\left\langle\langle\mathbf{h}|y\rangle^2\right\rangle_y &= \sum_{x,x'=1}^{n} h_xh_{x'}p^{2t}(x'|x)p(x)\\
&= \langle h_xh_{x'}\rangle_{2t}. \qquad (4.18)\\
\sigma^2(\langle\mathbf{h}|y\rangle) &= \left\langle\langle\mathbf{h}|y\rangle^2\right\rangle_y - \langle\mathbf{h}\rangle^2\\
&= \langle h_xh_{x'}\rangle_{2t} - \langle\mathbf{h}\rangle^2. \qquad (4.19)\\
\iota &= \frac{1}{2}\frac{\langle h_xh_{x'}\rangle_{2t} - \langle\mathbf{h}\rangle^2}{1-\langle\mathbf{h}\rangle^2}. \qquad (4.20)
\end{aligned}$$

B. Fast-mixing graphs

When diffusion on a graph reaches its well-mixed limit in short times, we have $\mathbf{G}^{2t} \approx \mathbf{I} - 2t\mathbf{\Delta P}^{-1}$. Thus, for a K-partition of a graph

$$\begin{aligned}
\mathbf{Q}(\mathbf{G}^{2t}\mathbf{P})\mathbf{Q^T} &\approx \mathbf{Q}(\mathbf{P} - 2t\mathbf{\Delta})\mathbf{Q^T}\\
&= \mathbf{QPQ^T} - 2t\mathbf{Q\Delta Q^T}. \qquad (4.21)
\end{aligned}$$

For bisections, the short-time approximation of $\langle h_x h_{x'} \rangle_{2t}$ can be written as

$$\begin{aligned} \langle h_x h_{x'} \rangle_{2t} &= \sum_{x,x'=1}^{n} h_{x'} p^{2t}(x', x) h_x \\ &= \mathbf{h^T G^{2t} P h} \\ &\approx \mathbf{h^T} (I - 2t\mathbf{\Delta P}^{-1})\mathbf{P h} \\ &= \mathbf{h^T P h} - 2t\mathbf{h^T \Delta h} \\ &= 1 - 2t\mathbf{h^T \Delta h}. \end{aligned} \tag{4.22}$$

For fast-mixing graphs, the long-time and short-time approximations for $I[Y;Z]$ and $\langle h_x h_{x'} \rangle_{2t}$, respectively, hold simultaneously.

$$\begin{aligned} I[Y;Z] \approx \quad \iota \quad &\approx \left(\tfrac{1}{2} - t\tfrac{\mathbf{h}^T \mathbf{\Delta h}}{1 - \langle \mathbf{h} \rangle^2}\right) \\ \Rightarrow \tfrac{dI[Y;Z]}{dt} \approx \tfrac{d\iota}{dt} &\propto \begin{cases} \mathcal{A} & ; \; p(x) \propto 1 \\ \mathcal{N} & ; \; p(x) \propto d_x. \end{cases} \end{aligned} \tag{4.23}$$

We have shown analytically that, for fast mixing graphs, the heuristics introduced by Shi and Malik are proportional to the rate of loss of relevance information. The error incurred in the approximations $I[Y;Z] \approx \iota$ and $\langle h_x h_{x'} \rangle_{2t} \approx 1 - 2t\mathbf{h^T \Delta h}$ can be defined as

$$\mathcal{E}_0(t) = \left| \frac{\langle h_x h_{x'} \rangle_{2t} - (1 - 2t\mathbf{h^T \Delta h})}{\langle h_x h_{x'} \rangle_{2t}} \right| \tag{4.24}$$

$$\mathcal{E}_1(t) = \left| \frac{I[Y;Z](t) - \iota(t)}{I[Y;Z](t)} \right|. \tag{4.25}$$

5. NUMERICAL EXPERIMENTS

The validity of the two approximations can be seen in a typical plot of $\mathcal{E}_1(t)$ and $\mathcal{E}_0(t)$ as a function of normalized diffusion time $\tilde{t} = t/\tau$, for the two different choices of prior distributions over the nodes. $\mathcal{E}_1$, as seen in Fig. 1, is often found to be non-monotonic and sometimes exhibits oscillations. This suggests defining $\mathcal{E}_\infty$, a modified monotonic '$\mathcal{E}_1$':

$$\mathcal{E}_\infty(t) \equiv \max_{t' \geq t} \mathcal{E}_1(t'). \tag{5.1}$$

We don't need to define a monotonic form for $\mathcal{E}_0$ since this error is always found to be monotonically increasing in time.

By fast-mixing graphs, we mean graphs which become well-mixed in short times, i.e. graphs for which both the long-time and short-time approximations hold simultaneously within a certain range of time $\tilde{t}^*_- \leq \tilde{t} \leq \tilde{t}^*_+$, as illustrated in Fig. 1, where we define

$$\mathcal{E}(t) = \max(\mathcal{E}_\infty(t), \mathcal{E}_0(t)) \tag{5.2}$$

$$\mathcal{E}^* = \min_t \mathcal{E}(t) \tag{5.3}$$

$$\tilde{t}^*_- = \min(\arg\min_{\tilde{t}} \mathcal{E}(\tilde{t})) \tag{5.4}$$

$$\tilde{t}^*_+ = \max(\arg\min_{\tilde{t}} \mathcal{E}(\tilde{t})). \tag{5.5}$$

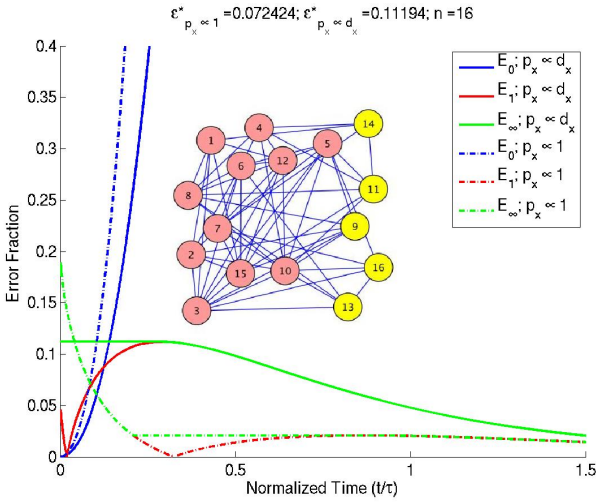

FIG. 1: $\mathcal{E}_1$ and $\mathcal{E}_0$ vs normalized diffusion time for two choices of priors over the graph nodes. $\mathcal{E}_1$ (red) typically tends to have a non-monotonic behavior which motivates defining a monotonic $\mathcal{E}_\infty$ (green).

Note that the use of $\mathcal{E}_\infty$ instead of $\mathcal{E}_1$ over-estimates the value of $\mathcal{E}^*$; the $\mathcal{E}^*$'s calculated is an upper bound.

Graphs were drawn randomly from a Stochastic Block Model (SBM) distribution [10], with block cardinality 2, to analyze the distribution of $\mathcal{E}^*$, $\tilde{t}^*_-$ and $\tilde{t}^*_+$. As is commonly done in community detection [11], for a graph of n nodes, the average degree per node is fixed at $n/4$ for graphs drawn from the SBM distribution: two nodes are connected with probability p_+ if they belong to the same block, but with probability $p_- < p_+$, if they belong to different blocks. The two probabilities are, thus, constrained by the relation

$$p_+ \left(\frac{n}{2} - 1\right) + p_- \left(\frac{n}{2}\right) = \frac{n}{4} \tag{5.6}$$

leaving only one free parameter p_- that tunes the 'modularity' of graphs in the distribution. Starting with a graph drawn from a distribution specified by a p_- value and specifying an initial cluster assignment as given by the SBM distribution, we make local moves — adding or deleting an edge in the graph and/or reassigning a node's cluster label — and search exhaustively over this moveset for local minima of $\mathcal{E}^*$. Fig. 2 compares the values of $\mathcal{E}^*$ and $\{\tilde{t}^*_-, \tilde{t}^*_+\}$ for graphs obtained in this systematic search, starting with a graph drawn from a distribution with $p_- = 0.02$ and $n = \{16, 32, 64\}$. We note that the scatter plots for graphs of different sizes collapse on one another when $\mathcal{E}^*$ is plotted against normalized time, confirming the Fiedler value $1/\tau$ to be an appropriate characteristic diffusion time-scale as used in [6]. A plot of $\mathcal{E}^*$ against actual diffusion time shows that the scatter plots of graphs of different sizes no longer collapse

Having shown analytically that for fast mixing graphs, the regularized mincut is approximately the rate of loss of relevance information, it would be instructive to compare

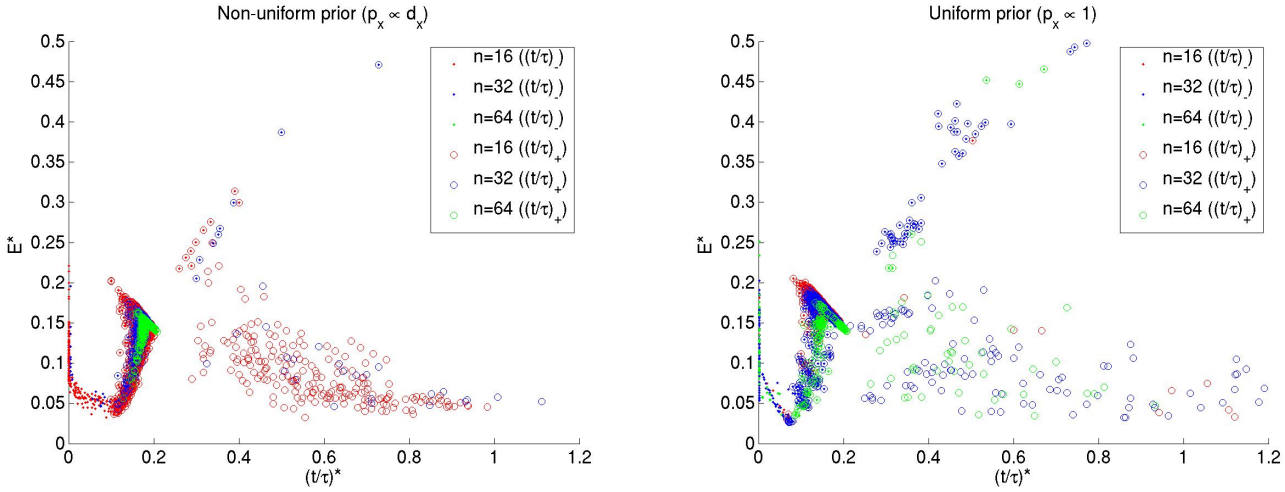

FIG. 2: $\mathcal{E}^*$ vs $\tilde{t}^*$ for graphs of different sizes and different prior distributions over the graph nodes. In the above plot, $\tilde{t}^*_-$ and $\tilde{t}^*_+$ are represented by $\cdot$ and $\circ$, respectively.

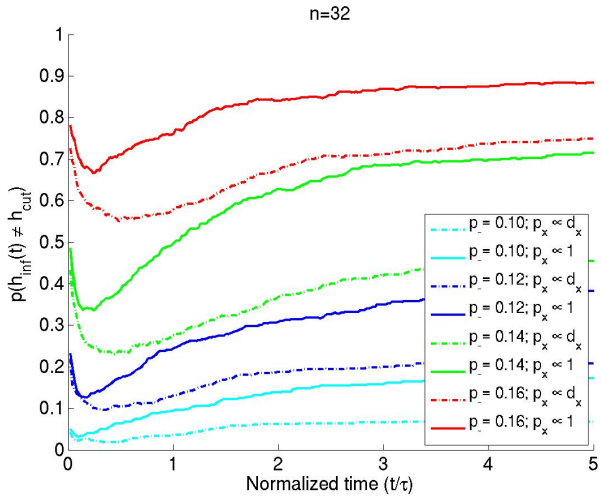

FIG. 3: $p(\mathbf{h}_{\rm inf}(t) \neq \mathbf{h}_{\rm cut})$ vs normalized diffusion time, averaged over 500 graphs drawn from a distribution parameterized by a given p_- value, is plotted for different graph distributions

the actual partitions that optimize these goals. Graphs of size $n = 32$ were drawn from the SBM distribution with $p_- = \{0.1, 0.12, 0.14, 0.16\}$. Starting with an equal-sized partition specified by the model itself, we performed iterative coordinate descent to search (independently) for the partition that minimized the regularized cut ($\mathbf{h}_{\rm cut}$) and one that minimized the relevance information ($\mathbf{h}_{\rm inf}(t)$); i.e. we reassigned each node's cluster label and searched for the reassignment that gave the new lowest value for the cost function being optimized. Plots comparing the partitions $\mathbf{h}_{\rm inf}(t)$ and $\mathbf{h}_{\rm cut}$, learnt by optimizing the two goals (averaged over 500 graphs drawn from each distribution), are shown in Fig. 3.

6. CONCLUDING REMARKS

We have shown that the normalized cut and average cut, introduced by Shi and Malik as useful heuristics to be minimized when partitioning graphs, are well approximated by the rate of loss of predictive information for fast-mixing graphs. Deriving these cut-based cost functions from rate-distortion theory gives them a more principled setting, makes them interpretable, and facilitates generalization to appropriate cut-based cost functions in new problem settings. We have also shown (see Fig. 2) that the inverse Fiedler value is an appropriate normalization for diffusion time, justifying its use in [6] to capture long-time behaviors on the network.

Absent from this manuscript is a discussion of how not to overpartition a graph, i.e. a criterion for selecting K. It is hoped that by showing how these heuristics can be derived from a more general problem setting, lessons learnt by investigating stablilty, cross-validation or other approaches may benefit those using min-cut based approaches as well. Similarily, by showing how these heuristics approximate costs functions from a separate optimization problem, it is hoped that algorithms employed for rate distortion theory, e.g. Blahut Arimoto, maybe be brought to bear on min-cut minimization.

APPENDIX

Using the definition of $\mathbf{\Delta}$, for any general vector $\mathbf{f}$ over the graph nodes, we have

$$\begin{aligned}\mathbf{f^T \Delta f} &= \mathbf{f^T D f} - \mathbf{f^T A f} \\ &= \sum_x d_x f_x^2 - \sum_{x,y=1}^{n} f_x f_y A_{xy}\end{aligned}$$

$$
\begin{aligned}
&= \sum_x \left(\sum_{y=1}^{n} A_{xy} \right) f_x^2 - \sum_{x,y=1}^{n} f_x f_y A_{xy} \\
&= \frac{1}{2} \left(\sum_{x,y=1}^{n} f_x^2 A_{xy} - 2 \sum_{x,y=1}^{n} f_x f_y A_{xy} + \sum_{x,y=1}^{n} f_y^2 A_{xy} \right) \\
&= \frac{1}{2} \sum_{x,y=1}^{n} A_{xy} \left(f_x - f_y \right)^2 . \qquad \text{(A.1)}
\end{aligned}
$$

Now, when $\mathbf{f} = \mathbf{h}$, we have

$$
\begin{aligned}
\mathbf{h^T \Delta h} &= \frac{1}{2} \sum_{h_x \times h_y = -1} 4 A_{xy} \\
&= 4 \times c. \qquad \text{(A.2)}
\end{aligned}
$$

The factor $\frac{1}{2}$ disappears because summation over all nodes counts each adjacent pair of nodes twice.

Using the definitions of $\mathcal{A}$ and $\mathcal{N}$, we have

$$
\begin{aligned}
\mathcal{A} &= c \times \left(\frac{1}{\sum_{h_x=+1} 1} + \frac{1}{\sum_{h_x=-1} 1} \right) \\
&= c \times \left(\frac{1}{\sum_x \left(\frac{1+h_x}{2} \right)} + \frac{1}{\sum_x \left(\frac{1-h_x}{2} \right)} \right) \\
&= 2c \times \left(\frac{\sum_x (1 - h_x + 1 + h_x)}{\sum_x (1 + h_x) \sum_x (1 - h_x)} \right) \\
&= 2c \times \left(\frac{2n}{(n + \sum_x h_x)(n - \sum_x h_x)} \right) \\
&= 2c \times \left(\frac{2}{n(1 + \bar{\mathbf{h}})(1 - \bar{\mathbf{h}})} \right) \\
&= \frac{4}{n} \frac{c}{1 - \bar{\mathbf{h}}^2}. \qquad \text{(A.3)}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{N} &= c \times \left(\frac{1}{\sum_{h_x=+1} d_x} + \frac{1}{\sum_{h_x=-1} d_x} \right) \\
&= c \times \left(\frac{1}{\sum_x d_x \left(\frac{1+h_x}{2} \right)} + \frac{1}{\sum_x d_x \left(\frac{1-h_x}{2} \right)} \right) \\
&= 2c \times \left(\frac{\sum_x d_x (1 - h_x + 1 + h_x)}{\sum_x \left(d_x (1 + h_x) \right) \sum_x \left(d_x (1 - h_x) \right)} \right) \\
&= 2c \times \left(\frac{4m}{(2m + \sum_x h_x d_x)(2m - \sum_x h_x d_x)} \right) \\
&= 2c \times \left(\frac{1}{m(1 + \langle \mathbf{h} \rangle)(1 - \langle \mathbf{h} \rangle)} \right) \\
&= \frac{2}{m} \frac{c}{1 - \langle \mathbf{h} \rangle^2}. \qquad \text{(A.4)}
\end{aligned}
$$

[1] J Shi and J Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jan 2000.

[2] N Tishby, F C Pereira, and W Bialek. The information bottleneck method. *arXiv preprint physics*, Jan 2000.

[3] N Tishby and N Slonim. Data clustering by markovian relaxation and the information bottleneck method. *Advances in Neural Information Processing Systems*, Jan 2000.

[4] C Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), Jan 2001.

[5] N Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2002.

[6] E Ziv, M Middendorf, and C H Wiggins. Information-theoretic approach to network modularity. *Physical Review E*, Jan 2005.

[7] U von Luxburg. A tutorial on spectral clustering. *arXiv*, cs.DS, Nov 2007.

[8] M Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 1973.

[9] Dorothea Wagner and Frank Wagner. Between min cut and graph bisection. In *MFCS '93: Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science*, pages 744–750, London, UK, 1993. Springer-Verlag.

[10] P W Holland and S Leinhardt. Local structure in social networks. *Sociological Methodology*, Jan 1976.

[11] L Danon, A Diaz-Guilera, J Duch, and A Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, Jan 2005.

[12] We chose 16-node graphs so the network and its partitions could be parsed visually with ease.

[13] We use the shorthand $x \sim y$ to mean x is adjacent to y.

[14] Strictly speaking, any diagonal matrix $\mathbf{P}$ that we specify determines the steady-state distribution. Since we are modeling the distribution of random walkers at statistical equilibrium, we always use this distribution as our initial or prior distribution.